Disguised Faces in the Wild 2019

Maneet Singh, Mohit Chawla, Richa Singh, Mayank Vatsa
IIIT-Delhi, India
{maneets, mohit17028, rsingh, mayank}@iiitd.ac.in

Rama Chellappa
University of Maryland, College Park, USA
rama@umiacs.umd.edu

Abstract

Disguised face recognition has wide-spread applicability in scenarios such as law enforcement, surveillance, and access control. Disguise accessories such as sunglasses, masks, scarves, or make-up modify or occlude different facial regions which makes face recognition a challenging task. In order to understand and benchmark the state-of-the-art on face recognition in the presence of disguise variations, the Disguised Faces in the Wild 2019 (DFW2019) competition has been organized. This paper summarizes the outcome of the competition in terms of the dataset used for evaluation, a brief review of the algorithms employed by the participants for this task, and the results obtained. The DFW2019 dataset has been released with four evaluation protocols and baseline results obtained from two deep learning-based state-of-the-art face recognition models. The DFW2019 dataset has also been analyzed with respect to degrees of difficulty: (i) easy, (ii) medium, and (iii) hard. The dataset has been released as part of the International Workshop on Disguised Faces in the Wild at International Conference on Computer Vision (ICCV), 2019.

1. Introduction

Automated face recognition systems often encounter the challenge of matching a face image captured in constrained settings (termed as *gallery*) with a disguised face image captured in an unconstrained environment (termed as *probe*). In unconstrained settings, the use of disguise accessories such as hats, scarves, sunglasses, helmets, head-bands, veils, turbans, or masks often result in occlusion of different face parts [5]. On the other hand, heavy make-up or external procedures such as plastic surgery can result in modifications to the face shape, texture, and color [2, 15]. Figure 1 presents sample face images of three subjects demonstrating the effects of using different disguise accessories or heavy make-up. The vast variations observed between the gallery and the probe images render disguised face recognition a challenging task.

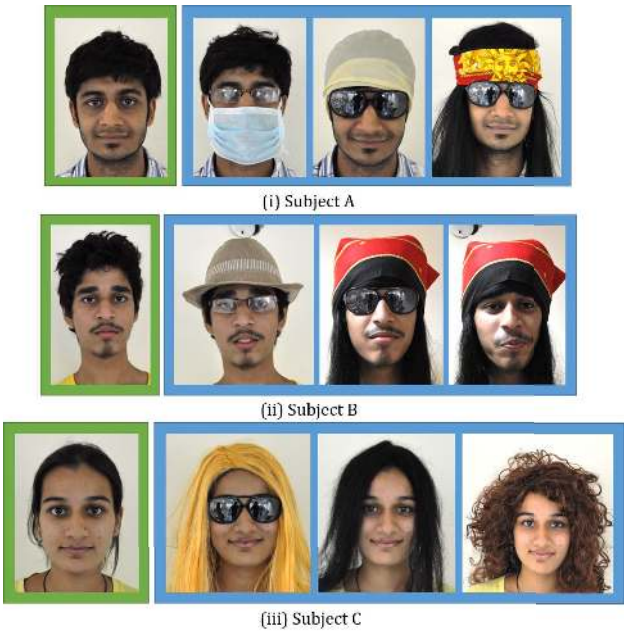


Figure 1: Sample images of three subjects demonstrating variations due to different disguise accessories. A face recognition system is required to match the *gallery* images (green box) with the *probe* images (blue box). Images are taken from the IIIT-Delhi Disguise dataset [5].

Often, disguise accessories such as hats, sunglasses, or scarves are used sub-consciously, without any real intent of occluding the face image. Coupled with scenarios where disguise accessories are used intentionally to mask one's identity, the task of disguised face recognition demands dedicated attention. Owing to the easy availability and wide usage of such accessories, disguised face recognition has large applicability in scenarios such as mobile phone or lap-

Table 1: Statistics of the DFW2019 dataset.

Image Variation	Number of	
	Subjects	Images
Bridal	100	200
Plastic surgery	250	500
Other	250	3140
Total	600	3840

top authentication via face unlock tools, automated facial tagging on social media, school attendance systems, and smart advertisements. Law enforcement applications such as surveillance, access control, and criminal identification can also benefit from a system robust to disguised faces.

Despite the wide-scale applicability of disguised face recognition, the problem has received limited attention from the research community. In the literature, most of the techniques have focused on disguised face recognition in constrained settings with limited disguise accessories [10, 12, 13, 16]. In 2016, the Disguise and Makeup dataset [19] was released, which contains disguised face images from publicly accessible websites. Recently, in 2018, the Disguised Faces in Wild dataset (referred to as DFW2018 dataset) [9, 14] was released as part of the International Workshop on Disguised Faces in the Wild, held in conjunction with the International Conference on Computer Vision and Pattern Recognition (CVPR), 2018. To the best of our knowledge, the DFW2018 dataset was a first-of-its-kind dataset capturing unconstrained variations across a wide spectra of disguise accessories and make-up, with the presence of *impersonators* for each subject. This research builds upon the DFW2018 dataset and presents the DFW2019 dataset. The dataset contains 3840 face images of 600 subjects, having variations across disguise accessories, bridal make-up, and plastic surgery. The dataset was released in the DFW2019 competition, as part of the International Workshop on Disguised Faces in the Wild, held in conjunction with the International Conference on Computer Vision (ICCV), 2019. This research summarizes the DFW2019 competition in terms of the DFW2019 dataset along with its four benchmark protocols, performance of the submissions, and the baseline results.

2. Disguised Faces in the Wild 2019 Dataset

The Disguised Faces in the Wild 2019 (DFW2019) dataset contains 3840 face images of 600 subjects. The images are collected from the Internet using relevant keywords from search engines, thereby demonstrating variability in terms of pose, illumination, resolution, acquisition mode, and disguise accessories. Other than images with external accessories such as hats, caps, beard, and sunglasses, the DFW2019 dataset also contains a subset of images having variations due to plastic surgery and bridal make-up. Broadly, the DFW2019 dataset contains two types of images: (i) subjects having before-after images with variations due to plastic surgery or bridal make-up, and (ii) subjects having unconstrained disguise variations due to occlusions or make-up, along with a *normal*, *validation*, and multiple *impersonator* images. Table 1 presents the statistics of the DFW2019 dataset. The dataset contains:

Figure 2: Sample impersonator pairs created from the IIIT-Delhi Disguise dataset [5]. An individual can often use disguise accessories to impersonate another individual.

- 200 **bridal** images of 100 subjects, where each subject has two images corresponding to before and after applying bridal make-up,
- 500 **plastic surgery** images of 250 subjects, where each subject has two images corresponding to before and after the plastic surgery procedure,
- 3140 images of 250 subjects, where each subject contains a **validation** and a **normal** image, which corresponds to frontal and non-disguised high resolution images having good illumination. Each subject also contains a set of **disguised** images and a set of **impersonator** images (different subjects which appear to intentionally/unintentionally impersonate the subject). Figure 2 presents sample impersonator pairs.

The dataset will be made available for the research community[1]. Four protocols have also been defined for evaluations on the DFW2019 dataset. The following subsection elaborates upon each of the protocols.

2.1. Protocols for Evaluation

Four verification protocols have been presented for evaluating face recognition algorithms on the DFW2019 dataset. Continuing from the DFW2018 competition [14], two protocols are: (i) Impersonation and (ii) Obfuscation, while the remaining two correspond to (iii) Plastic Surgery and (iv) Overall. The following paragraphs present each protocol in detail, along with the description of genuine and imposter sets.

Protocol 1 - Impersonation: This protocol aims to assess a face recognition system under the effect of impersonation.

[1] http://iab-rubric.org/resources.html

Table 2: Baseline results on the DFW2019 dataset. GAR is reported for the specified FAR values.

Protocol	Model	0.1% FAR	0.01%FAR
P-1	ResNet-50	47.6	38.4
	LightCNN-29v2	74.4	51.2
P-2	ResNet-50	35.3	16.4
	LightCNN-29v2	55.5	36.9
P-3	ResNet-50	46.4	22.4
	LightCNN-29v2	69.2	47.2
P-4	ResNet-50	35.9	16.8
	LightCNN-29v2	55.7	36.5

Here, the genuine set consists of the normal-validation image pair of the same subject, and the imposter set consists of normal-impersonator pair, disguise-impersonator pair, and validation-impersonator pair of the same subject. In this protocol, there exist 250 genuine and 7,431 imposter pairs.

Protocol 2 - Obfuscation: This protocol focuses on evaluating a face recognition system under intentional or unintentional disguise variations of genuine users. Here, the genuine set corresponds to the normal-disguise, validation-disguise, and disguise_1-disguise_2 image pairs of the same subject, along with the before-after bridal make-up images. The imposter set contains cross-subject pairs, where the disguised, normal, and validation images of one subject are paired with the disguised, normal, and validation images of another subject. Moreover, cross-subject before-after pairs for the bridal make-up set also constitute the imposter set. In total, this protocol contains 10,267 genuine and 2,802,011 imposter pairs.

Protocol 3 - Plastic Surgery: This protocol is specifically targeted towards evaluating a face recognition system against changes in facial features due to plastic surgery. Here, the before-after images of subjects who have undergone plastic surgery are utilized. The genuine set (250 pairs) contains the before-after images of the same subject, while the imposter set (124,500 pairs) contains cross-subject before-after images.

Protocol 4 - Overall: The overall protocol attempts to evaluate a face recognition system on the entire DFW2019 dataset. Here, the genuine set contains a combination of all the images in the genuine sets of Protocols 1-3. That is, the genuine set contains the normal-validation (Protocol-1), validation-disguise, normal-disguise, disguise_1-disguise_2, before-after bridal make-up (Protocol-2), and before-after plastic surgery (Protocol-3) image pairs. The imposter set also contains a combination of the imposter pairs across Protocols 1-3. That is, the imposter set contains normal-impersonator, disguise-impersonator, validation-impersonator (Protocol-1), cross-subject imposters, cross-subject before-after bridal make-up (Protocol-2), and cross-subject before-after plastic surgery (Protocol-3) pairs.

3. Baseline Results

For all the protocols, baseline results have been computed using two pre-trained state-of-the-art deep learning based face recognition models. ResNet-50[2] [7] (pre-trained on the large-scale VGG-Face2 [1] and MS-Celeb-1M [6] datasets) and LightCNN-29v2[3] [20] (pre-trained on the large-scale CASIA-WebFace [21] and MS-Celeb-1M [6] datasets) have been used for evaluation. Pre-trained models were used as is, without any additional training. Detected and cropped face images were provided to the network, followed by feature extraction, and Cosine similarity based classification. Face detection was performed using the Tiny Face detector [8], followed by manual detection of the false negative faces. The extracted embeddings were of dimension 2048 and 256 for ResNet-50 and LightCNN-29v2, respectively. Genuine Acceptance Rate (GAR) is reported for fixed False Acceptance Rates (FARs), which form the baselines for the DFW2019 dataset.

Table 2 presents the baseline results obtained for the DFW2019 dataset using the two networks: Resnet-50 and LightCNN-29v2. Results have been tabulated for two FARs: 0.1% and 0.01% for all the protocols (protocol 1-4). LightCNN-29v2 consistently outperforms the ResNet-50 model by achieving improved verification performance across all protocols and FARs.

4. Disguised Faces in the Wild 2019 Competition

The DFW2019 competition[4] was held in conjunction with the *International Workshop on Disguised Faces in the Wild* at the International Conference on Computer Vision (ICCV), 2019. Participants had to develop a face recognition model which is evaluated on the DFW2019 dataset.

Anonymized DFW2019 dataset was provided to the participants as the test set, and evaluation is performed on all four protocols. The training and testing partitions of the DFW2018 dataset [14] were also provided as the training and validation partition, respectively, for the competition. The DFW2019 dataset will be made publicly available for the research community. We believe that the DFW2019 dataset can help in enhancing the recognition performance for disguised faces, thereby improving the robustness of face recognition algorithms.

4.1. DFW2019 Competition: Submissions

The DFW2019 competition received over 100 registrations and 11 submissions from all over the world. Table 3 summarizes the affiliation of the different submissions received as part of this competition. Each submission is

[2]https://github.com/cydonia999/VGGFace2-pytorch
[3]https://github.com/AlfredXiangWu/LightCNN
[4]http://iab-rubric.org/DFW/2019Competition.html

Table 3: List of teams who participated in the DFW2019 competition.

Algorithm	Team	Institution
A-1	ArcFace	Imperial College London
A-2	ArcFaceInter	Imperial College London
A-3	ArcFaceIntra	Imperial College London
A-4	ArcFaceIntraInter	Imperial College London
A-5	FakeFace	ITMO University
A-6	FakeFacev2	ITMO University
A-7	FEBNet	Indian Institute of Technology, Madras
A-8	LightCNNDFW	Anonymous
A-9	Mozart	Tech5.ai
A-10	SEBNet	Indian Institute of Technology, Madras
A-11	XuXu	Tech5.ai

described in detail as follows:

(i) ArcFace: A team from the Imperial College London proposed using ArcFace [3] (Additive Angular Margin Loss) for recognizing disguised faces in the wild. The model incorporates a margin in the popularly used Softmax loss for deep learning based Convolutional Neural Networks. Facial co-ordinates are computed using the RetinaFace model [4].

(ii) ArcFaceInter: Submitted by a team from the Imperial College London, ArcFaceInter incorporates an additional term for enhancing the inter-class distance in the ArcFace [3] model. RetinaFace [4] is used for computing the facial co-ordinates and geometric alignment of images.

(iii) ArcFaceIntra: ArcFaceIntra incorporates an intra-class penalty to enhance class compactness into the ArcFace model. Submitted by a team from the Imperial College London, features are extracted from the ArcFaceIntra model for faces detected and aligned via the RetinaFace model [4].

(iv) ArcFaceIntraInter: ArcFaceIntraInter models both inter-class and intra-class variations during feature learning. Submitted by a team from the Imperial College London, ArcFaceIntraInter incorporates two additional terms in the ArcFace model [3] for increasing the inter-class distance and reducing the intra-class variations. Face detection and alignment is performed using the RetinaFace [4] model.

(v) FakeFace: Submitted by a team from the ITMO University, Russia, faces are detected and aligned with RetinaFace [4] and cropped to 112×112. A deep learning network is trained using the MS-Celeb-1M dataset [6] and the ArcFace loss [3]. The model is fine-tuned with Doppleganger Mining [17], Auxillary Embeddings [18], Embeddings Interpolations, and Priority Lists. An ensemble of three such networks is used for feature extraction.

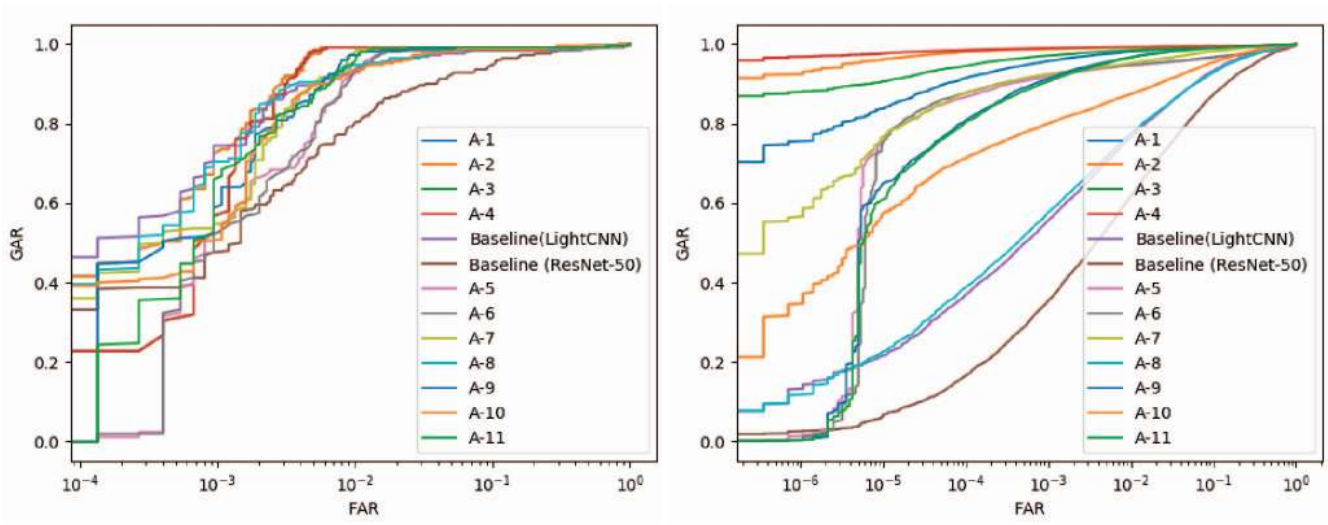
(vi) FakeFacev2: Submitted by a team from the ITMO University, Russia, FakeFacev2 uses a combination of RetinaFace [4] and ArcFace [3] as backbone for recognizing disguised faces in the wild. Fine-tuning is performed on the MS-Celeb-1M dataset [6] with Doppleganger Mining [17], Auxillary Embeddings [18], Embeddings Interpolations, and Priority Lists. Evaluation is performed using an ensemble of three such networks.

(vii) FEBNet: A team from the Indian Institute of Technology, Madras proposed the FEBNet model. Detected faces provided with the dataset are used with an ensemble of SE-ResNet-50 (pre-trained on the MS-Celeb-1M dataset [6]) and Inception-ResNet-v1 (pre-trained on the VGGFace2 dataset [1]). Fine-tuning is performed using a combination of identity loss, triplet loss, and category loss. Decision is taken via score-level fusion and a re-ranking approach.

(viii) LightCNNDFW: A pre-trained LightCNN-29v2 [20] network has been fine-tuned in a Siamese manner. Binary cross-entropy loss is applied on the extracted features. Detected faces provided with the dataset are used, along with the *five-crop* data augmentation technique.

(ix) Mozart: Submitted by a team from Tech5.ai, Mozart uses the detected faces provided with the DFW2019 dataset. An ensemble of different ResNet models is used for feature extraction, followed by matching via the l_2-distance.

(x) SEBNet: SEBNet has been submitted by a team from the Indian Institute of Technology, Madras and utilizes an ensemble of deep learning networks. Two networks: InceptionNet-v3 (pre-trained on the MS-Celeb-1M dataset [6]) and SE-ResNet-50 (pre-trained on the VGGFace2 dataset [1]) are fine-tuned on the DFW2018 dataset. The



(a) Protocol-1 (Impersonation)

(b) Protocol-2 (Obfuscation)

Figure 3: ROC curves on the DFW2019 dataset for Protocol-1 and Protocol-2.

Table 4: Verification accuracy (%) on the proposed DFW2019 dataset for the Impersonation protocol (Protocol-1). The table presents the performance of participants and the baseline results.

Algorithm	GAR	
	@0.1%FAR	@0.01%FAR
A-1	72.4	44.8
A-2	72.4	44.8
A-3	56.8	17.6
A-4	56.8	17.6
A-5	52.4	1.2
A-6	52.0	2.0
A-7	54.8	42.4
A-8	70.4	43.2
A-9	58.8	44.8
A-10	54.8	40.0
A-11	66.0	24.4
Baseline (LightCNN)	**74.4**	**51.2**
Baseline (ResNet-50)	47.6	38.4

Table 5: Verification accuracy (%) on the proposed DFW2019 dataset for the Obfuscation protocol (Protocol-2). The table summarizes the performance of participants and the baseline results.

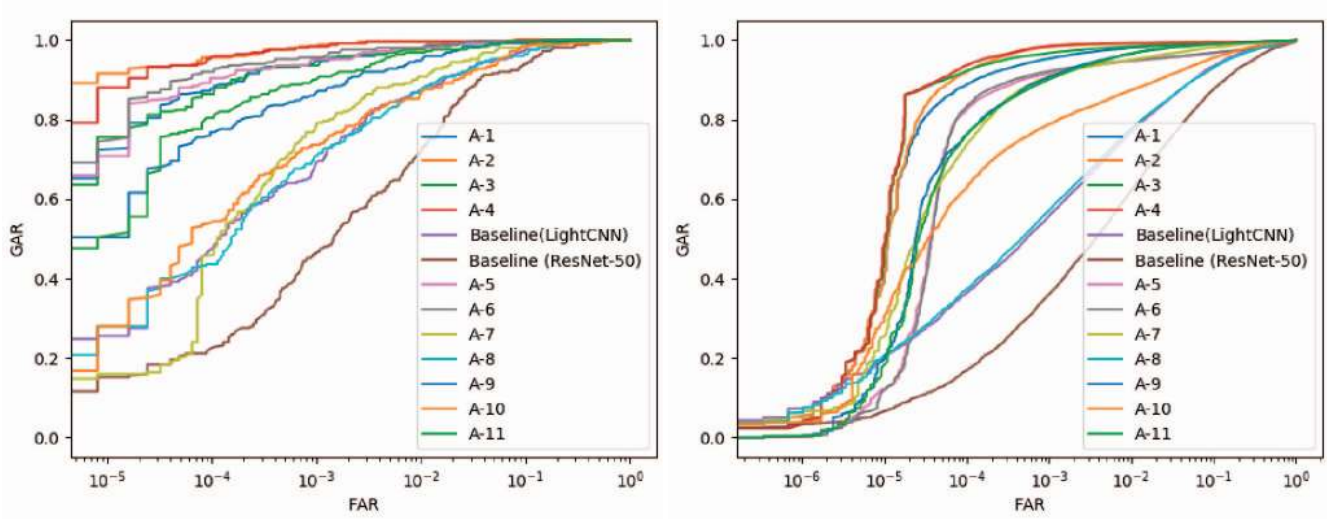
Algorithm	GAR	
	@0.1%FAR	@0.01%FAR
A-1	95.7	91.4
A-2	98.7	97.9
A-3	97.0	94.4
A-4	**98.9**	**98.4**
A-5	91.6	86.6
A-6	92.3	87.7
A-7	92.3	87.6
A-8	57.5	38.6
A-9	91.1	80.5
A-10	80.0	71.2
A-11	90.5	80.5
Baseline (LightCNN)	55.5	36.9
Baseline (ResNet-50)	35.3	16.4

trained networks are used for feature extraction, followed by Euclidean distance based score computation, score-level fusion, and a re-ranking algorithm.

(xi) XuXu: Submitted by a team from Tech5.ai, XuXu utilizes an ensemble of different ResNet models. The pipeline includes geometric alignment on the detected faces provided with the dataset, followed by feature extraction from the ensemble. Matching is performed using l_2-distance.

4.2. Results

For all the protocols, results are reported in the form of Genuine Acceptance Rate (GAR) for the specified False Acceptance Rates (FAR). Baseline results have been reported using the LightCNN-29v2 model [20] and the ResNet-50 model [7], with Cosine similarity based classification (Section 3). The following paragraphs elaborate upon the results obtained by for each protocol, including the submissions and the baseline results:



(a) Protocol-3 (Plastic Surgery)

(b) Protocol-4 (Overall)

Figure 4: ROC curves on the DFW2019 dataset for Protocol-3 and Protocol-4.

Table 6: Verification accuracy (%) for the Plastic Surgery protocol (Protocol-3). Results of the submissions and baseline performance computed using ResNet-50 and LightCNN-29v2 have been presented in the table.

Algorithm	GAR	
	@0.1%FAR	@0.01%FAR
A-1	94.8	87.6
A-2	**98.4**	**95.6**
A-3	93.6	86.4
A-4	**98.4**	**95.6**
A-5	95.2	90.4
A-6	95.6	92.0
A-7	78.8	47.6
A-8	70.8	43.6
A-9	86.8	76.8
A-10	73.6	54.0
A-11	90.0	81.2
Baseline (LightCNN)	69.2	47.2
Baseline (ResNet-50)	46.4	22.4

Table 7: Verification accuracy (%) for the Overall protocol (Protocol-4). The table presents the performance of the participants and baseline results computed using ResNet-50 and LightCNN-29v2.

Algorithm	GAR	
	@0.1%FAR	@0.01%FAR
A-1	95.2	88.6
A-2	98.3	92.0
A-3	96.7	92.1
A-4	**98.4**	**93.6**
A-5	91.4	82.2
A-6	92.1	83.1
A-7	90.7	73.6
A-8	57.1	37.4
A-9	90.7	76.1
A-10	78.8	62.8
A-11	90.0	76.0
Baseline (LightCNN)	55.7	36.5
Baseline (ResNet-50)	35.9	16.8

(i) Protocol-1 (Impersonation): Figure 3(a) contains the ROC curves for the baseline results and the submissions. Table 4 presents the GAR at 0.1% and 0.01% FAR for all the submissions. At both the FARs, the baseline performance of LightCNN-29v2 performs the best by achieving 74.4% and 51.2%, respectively. At both the FARs, A-1 (ArcFace) and A-2 (ArcFaceInter) perform second best with GARs of 72.4% and 44.8%, respectively. A drop of around 24% is observed between the verification performance at 0.1% and 0.01% FAR of LightCNN-29v2, suggesting the need for face recognition models to focus more on preventing impersonation based attacks.

(ii) Protocol-2 (Obfuscation): Figure 3(b) demonstrates the ROC curves for Protocol-2 (obfuscation), and Table 5 presents the GAR values at two specified FARs: 0.1% and 0.01%. A-4 (ArcFaceIntraInter) outperforms other techniques by reporting a GAR of 98.9% and 98.4% at 0.1% and 0.01% FAR, respectively. The second and third best perfor-

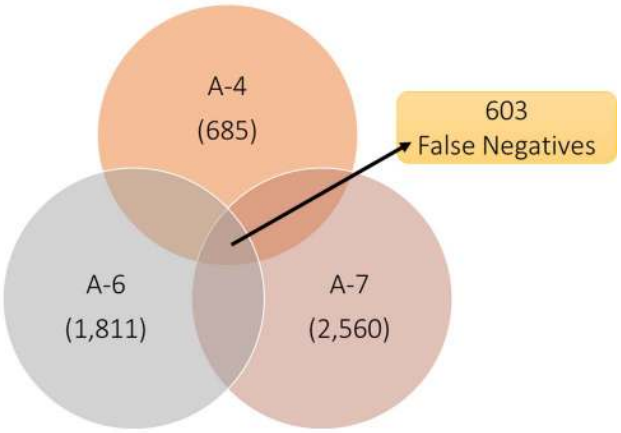


Figure 5: Venn diagram demonstrating the number of misclassifications of the genuine pairs by the top-3 teams (A-4: ArcFaceIntraInter, A-6: FakeFacev2, A-7:Mozart) at 0.01% FAR. The common region (603 samples) is a subset of the *hard* samples which were mis-classified by all algorithms.

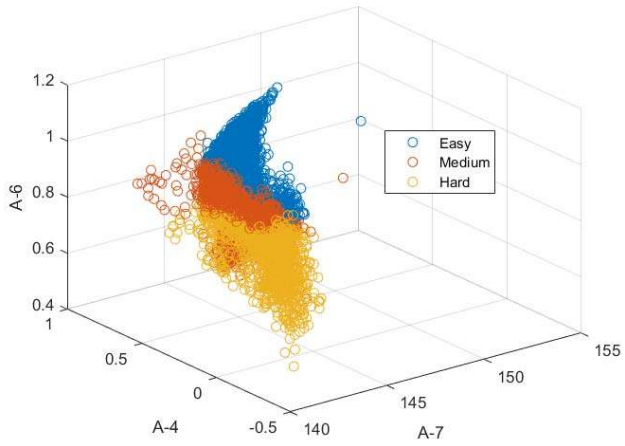


Figure 6: Scatter plot of the scores obtained by the top-3 teams for the Easy, Medium, and Hard pairs of the DFW2019 dataset.

mance are also obtained by variants of the ArcFace model. In Protocol-2 the variations observed between the GAR at 0.1% and 0.01% is less than that obtained in Protocol-2. The improved GARs at lower FARs further suggest that deep learning based face recognition models are able to handle variations due to obfuscation better, that is, scenarios where a genuine user attempts to obfuscate their identity by means of an external accessory.

(iii) Protocol-3 (Plastic Surgery): Figure 4(a) presents the ROC curves for Protocol-3, that is, variations brought in the face due to the plastic surgery procedure. Table 6 also presents the GAR values obtained at the specified FARs of 0.1% and 0.01% for all the submissions and baseline results. Best performance of 98.4% and 95.6% is obtained via A-2 (ArcFaceInter) and A-4 (ArcFaceIntraInter) for 0.1% and 0.01% FAR, respectively. The second and third best performance are obtained by A-6 (FakeFacev2) and A-5 (FakeFace) submissions, wherein a difference of around 3% is observed at 0.1%FAR. High verification performance on both FARs demonstrate the effectiveness of the submissions for handling face recognition under variations due to plastic surgery.

(iv) Protocol-4 (Overall): Protocol-4 evaluates the performance of a face recognition system on the entire DFW2019 dataset. Figure 4(b) presents the ROC curves of the submissions and baseline results, and Table 7 presents the GAR values obtained at 0.1% and 0.01% FAR, respectively. A-4 (ArcFaceIntraInter) achieves the highest performance on both the FARs: 98.4% and 93.6% at 0.1% and 0.01% FAR, respectively. This is followed by A-2 (ArcFaceInter) and A-3 (ArcFaceIntra) on both the FARs.

Overall, the DFW2019 competition received 11 submissions, all of which utilized deep learning based pre-trained networks. It is our belief that the availability of networks pre-trained on large datasets facilitates discriminative feature extraction, resulting in high performance.

Table 8: Total *easy*, *medium*, and *hard* pairs at 0.01% FAR. *Easy* refers to the number of pairs correctly classified as TP (True Positive)/TN (True Negative). *Medium* refers to the number of pairs correctly classified as TP/TN by two algorithms, while *Hard* refers to the number of TP/TN pairs correctly classified by at most one algorithm.

	Genuine (TP)	**Imposter (TN)**	**Total**
Easy	7,743	2,933,312	2,941,055
Medium	1,595	445	2,040
Hard	1,429	185	1,614

5. DFW2019 Dataset: Easy, Medium, and Hard Pairs

Based on the degree of difficulty of verifying a pair of face images, the DFW2019 dataset is divided into three components: *easy*, *medium*, and *hard*. This section presents an analysis of the dataset along the above mentioned components. The *easy* partition contains those image pairs which are relatively easy to correctly verify by face recognition algorithms. On the other hand, the *hard* partition contains those image pairs which are harder to verify by face recognition algorithms. Division of the DFW2019 dataset in easy, medium, and hard categories is similar in concept to the partitioning of the DFW2018 dataset [14], as well as the Good, Bad, and Ugly components of the FRVT 2006

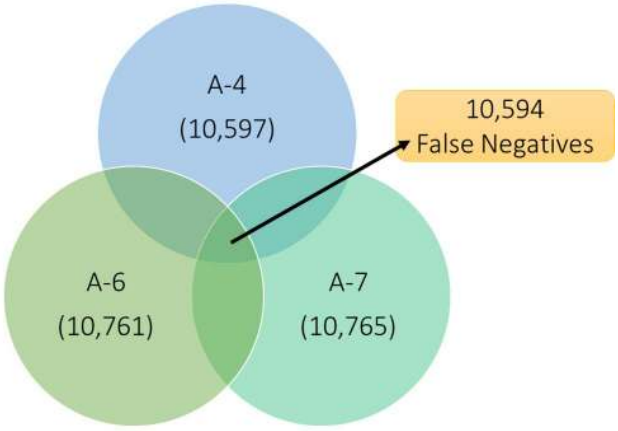


Figure 7: Venn diagram demonstrating the number of misclassifications of the genuine pairs (True Positive samples) by the top-3 teams (A-4: ArcFaceIntraInter, A-6: FakeFacev2, A-7:Mozart) for 0 False Positives. The common region (10,594 samples) corresponds to a subset of samples which were mis-classified by all algorithms.

competition dataset [11].

For the DFW2019 dataset, the results obtained by the top-3 teams for the Overall protocol (protocol-4) have been utilized to create the (i) *easy*, (ii) *medium*, and (iii) *hard* partition. As observed from Table 7, the top-3 teams correspond to: (i) A-4 (ArcFaceIntraInter), (ii) A-6 (FakeFacev2), and (iii) A-7 (Mozart). For the DFW2019 dataset, the *easy* partition corresponds to the image pairs correctly classified by all three algorithms. The *medium* partition contains pairs of face images which have been correctly classified by any two of the top-3 submitting teams, while the *hard* partition contains image pairs which have been classified correctly by any one algorithm, or have been incorrectly matched by all three algorithms. The partitioning of the DFW2019 dataset has been performed for both genuine and imposter pairs, and mutual exclusion has been ensured across the three partitions.

Table 8 presents the count of the easy, medium, and hard pairs for the DFW2019 dataset at 0.01%FAR. For the genuine set, 7,743 pairs belong to the *easy* category which were correctly matched by the top three teams. On the other hand, 1,429 pairs correspond to the genuine *hard* partition which were incorrectly classified by at least two of the top three teams (almost 14% of the entire genuine set). Figure 5 presents a Venn Diagram demonstrating the number of misclassifications of genuine pairs from the DFW2019 dataset. It can be observed that 603 pairs were mis-classified by all top-3 teams, which form a part of the hard partition for the DFW2019 dataset. In total, the medium and hard partitions correspond to 3,654 pairs of face images from the DFW2019 dataset. Figure 6 presents the scores obtained by the top-3 algorithms for the three partitions. Scores for the easy and hard sets of the DFW2019 dataset occupy opposite ends of the distribution, while scores corresponding to the medium partition are present in the middle.

In several law enforcement applications, face recognition systems are often required to operate under the strict threshold of 0% FAR. That is, no imposter pair should be incorrectly classified as a genuine pair (0 FAR) while correctly classifying the genuine set of images (high GAR). On the DFW2019 dataset, Figure 4(b) can be analyzed to observe very low performance at lower FARs for the overall protocol. Figure 7 presents a Venn Diagram for the number of incorrect classifications of the genuine set by the top-3 algorithms at 0% FAR. 10,594 pairs of images are mis-classified by all three algorithms, which corresponds to 98.39% of the total genuine samples. The reduced performance at lower FARs suggests the need for robust face recognition systems applicable to critical law enforcement applications. It is our belief that moving forward, face recognition algorithms should focus on further reducing the number of hard pairs, while achieving high accuracy on the easy partition.

6. Conclusions and Future Work

This research presents a novel Disguised Faces in the Wild 2019 (DFW2019) dataset, containing 3840 images of 600 subjects. All images are collected from the Internet via relevant keyword searches on different search engines, thereby demonstrating wide variations with respect to pose, illumination, lighting, resolution, capturing device, and disguise accessories. The DFW2019 dataset contains variations due to different disguise accessories, and before-after images for plastic surgery and bridal make-up. This research also presents four protocols and baseline results of two state-of-the-art deep learning based networks: LightCNN-29v2 [20] and ResNet-50 [7]. The four protocols used for evaluation correspond to: (i) Protocol-1 (Impersonation), (ii) Protocol-2 (Obfuscation), (iii) Protocol-3 (Plastic Surgery), and (iv) Protocol-4 (Overall). The dataset has been released as part of a competition held with the International Workshop on Disguised Faces in the Wild, in conjunction with the International Conference on Computer Vision (ICCV), 2019. This research also summarizes the performance of the 11 submissions received as part of the competition, and analysis has also been performed by partitioning the DFW2019 dataset into three components: (i) easy, (ii) medium, and (iii) hard. Performance of the top-3 teams from the DFW2019 competition has been analyzed to obtain the partitioning. It is our belief that the availability of the DFW2019 dataset will further facilitate the development of robust face recognition systems.

References

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 67–74, 2018.

[2] Antitza Dantcheva, Cunjian Chen, and Arun Ross. Can facial cosmetics affect the matching accuracy of face recognition systems? In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 391–398, 2012.

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[4] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019.

[5] Tejas I. Dhamecha, Richa Singh, Mayank Vatsa, and Ajay Kumar. Recognizing disguised faces: Human and machine evaluation. *PLOS ONE*, 9(7):1–16, 07 2014.

[6] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102, 2016.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[8] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 951–959, 2017.

[9] Vineet Kushwaha, Maneet Singh, Richa Singh, Mayank Vatsa, Nalini Ratha, and Rama Chellappa. Disguised faces in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2018.

[10] Billy YL Li, Ajmal S Mian, Wanquan Liu, and Aneesh Krishna. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *IEEE Workshop on Applications of Computer Vision*, pages 186–192, 2013.

[11] P. Jonathon Phillips, J. Ross Beveridge, Bruce A. Draper, Geof Givens, Alice J. O'Toole, David S. Bolme, Joseph Dunlop, Yui Man Lui, Hassan Sahibzada, and Samuel Weimer. An introduction to the good, the bad, the ugly face recognition challenge problem. In *Face and Gesture*, pages 346–353, 2011.

[12] Narayanan Ramanathan, Rama Chellappa, and AK Roy Chowdhury. Facial similarity across age, disguise, illumination and pose. In *International Conference on Image Processing*, volume 3, pages 1999–2002, 2004.

[13] Giulia Righi, Jessie J. Peissig, and Michael J. Tarr. Recognizing disguised faces. *Visual Cognition*, 20(2):143–169, 2012.

[14] Maneet Singh, Richa Singh, Mayank Vatsa, Nalini K. Ratha, and Rama Chellappa. Recognizing disguised faces in the wild. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(2):97–108, 2019.

[15] Richa Singh, Mayank Vatsa, Himanshu S. Bhatt, Samarth Bharadwaj, Afzel Noore, and Shahin S. Nooreyezdan. Plastic surgery: A new dimension to face recognition. *IEEE Transactions on Information Forensics and Security*, 5(3):441–448, 2010.

[16] Richa Singh, Mayank Vatsa, and Afzel Noore. Face recognition with disguise and single gallery images. *Image and Vision Computing*, 27(3):245 – 257, 2009.

[17] Evgeny Smirnov, Aleksandr Melnikov, Sergey Novoselov, Eugene Luckyanets, and Galina Lavrentyeva. Doppelganger mining for face representation learning. In *IEEE International Conference on Computer Vision Workshops*, pages 1916–1923, 2017.

[18] Evgeny Smirnov, Aleksandr Melnikov, Andrei Oleinik, Elizaveta Ivanova, Ilya Kalinovskiy, and Eugene Luckyanets. Hard example mining with auxiliary embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 37–46, 2018.

[19] Tsung Ying Wang and Ajay Kumar. Recognizing human faces under disguise and makeup. In *IEEE International Conference on Identity, Security and Behavior Analysis*, 2016.

[20] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[21] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014.